Chemometrics for analytical data mining in separation process design for recovery of artemisinin from Artemisia annua

Malwade, Chandrakant Ramkrishna; Qu, Haiyan; Rong, Ben-Guang; Christensen, Lars Porskjær

Published in:
Computer - Aided Chemical Engineering

Publication date:
2013

Document version
Final published version

Citation for pulished version (APA):
Malwade, C. R., Qu, H., Rong, B-G., & Christensen, L. P. (2013). Chemometrics for analytical data mining in separation process design for recovery of artemisinin from Artemisia annua. Computer - Aided Chemical Engineering, 32, 49-54.

Andrzej Kraslawski and Ilkka Turunen (Editors) Proceedings of the 23rd European Symposium on Computer Aided Process Engineering – ESCAPE 23, June 9-12, 2013, Lappeenranta, Finland

Chemometrics for analytical data mining in separation process design for recovery of artemisinin from *Artemisia annua*

Chandrakant R. Malwade, Haiyan Qu, Ben-Guang Rong*, Lars P. Christensen

Department of Chemical Engineering, Biotechnology and Environmental Technology, University of Southern Denmark, DK-5230, Odense M, Denmark; Tel. +45 6550 7481, email: bgr@kbm.sdu.dk

Abstract

In this work, to understand the separation behavior of the flash column chromatography during purification of artemisinin from the crude extract of *Artemisia annua*, multivariate data analysis technique PARAFAC (Parallel factorization) is used to mine the relevant chemical information from analytical chromatograms of 9 artemisinin containing fractions. The size of three way dataset obtained from chromatogram measurements in sample, retention time, and spectral mode is 9 × 1981 × 82. Prior to the application of PARAFAC, the dataset is preprocessed to remove baseline drift and peak misalignment caused by retention time shifts due to matrix effects. Due to the complicated nature of chromatograms, the preprocessed HPLC data were divided into intervals containing analytical signals and then PARAFAC modeling was performed on individual intervals. Loadings from the PARAFAC analysis provided pure elution profiles and pure UV spectra even for co-eluting peaks, thus enabling the identification of chromatographically unresolved components. Also, loadings were used to determine the number of components and their relative concentrations in the fractions containing artemisinin which are the most important information of the flash column performance.

Keywords: artemisinin, chemometrics, PARAFAC

1. Introduction

Designing an efficient separation process for recovery of natural products is a challenging task due to the presence of many unknown impurities and lack of fundamental process data. Therefore, rigorous use of process analytical technology (PAT) for qualitative as well as quantitative analysis of raw materials, process streams and final products during conceptual process design stage is inevitable. However, intensive use of PAT often generates enormous amount of data containing many variables which is often difficult to interpret and it is time consuming. Therefore, an efficient method is required to mine the relevant chemical information such as pure elution profiles, UV spectra, concentrations of desired compounds etc. from the vast data in order to expedite the process design for recovery of natural products.

Multivariate data analysis techniques such as principle component analysis (PCA), multi linear regression (MLS), partial least squares (PLS), and PARAFAC, a generalization of PCA extended to multi way arrays provides a suitable means of extracting useful information from vast pool of data (S. Wold, 1995; R. Bro, 1997). These techniques have been now a days widely used in different areas for rapid analysis of complicated data (P. Geladi *et al*., 2004). Recently, FDA has released draft guidance

for industry, in which it has strongly recommended use of PAT framework in manufacturing as well as process design stage for better understanding of processes (U.S. Food and Drug Administration, 2004). Chemometrics has been mentioned as an important tool combined with advanced process analyzers in this PAT framework.
The objective of the present work is to use multivariate data analysis technique PARAFAC as an aid to quicken the process design for recovery of the anti-malarial drug, artemisinin from dried leaves of the medicinal plant *Artemisia annua*. Our preliminary lab scale experiments have confirmed that it is feasible to combine chromatography and crystallization operations to get the target compound and it has been found that the chemical composition of fractions obtained by chromatographic separation is the key to obtain synergistic effect between these two operations (C. Malwade *et al*., 2012). Detailed analysis of the fractions containing artemisinin by HPLC equipped with a Diode Array Detector (DAD) is carried out in order to get insight into the chemical composition of fractions. However, the HPLC measurements produced a large array of data and it was very time consuming to look at single peaks from the crowded chromatograms and manually integrate them to determine the concentration. Therefore, advanced multivariate data analysis technique, PARAFAC is used to analyse chromatograms in order to determine the number of chemical components present in each fraction in addition to artemisinin, their relative concentration profiles and pure UV spectra.
The algorithms used in the present work are executed in MATLAB®2010b software and calculations are performed on IBM PC with Intel CORE i7 1.60 GHz processor and 4 GB of installed memory (RAM).

2. Experimental

2.1. Extraction of artemisinin from dried leaves of Artemisia annua

Artemisinin was extracted from dried leaves of *A. annua* with dichloromethane (DCM) using maceration technique. DCM was chosen as the solvent mainly due to the high solubility of artemisinin in it combined with its low boiling point that facilitates easy recovery of the solvent. The extraction procedure included immersion of 150 g of dried leaves containing 2.05% w/w artemisinin into 1.5 l of DCM at room temperature followed by filtration after 6 hrs. The procedure was repeated with 1 l of fresh DCM and the combined extract (2.5 l) was evaporated to obtain 12.5 g of crude extract.

2.2. Purification of crude extract by flash column chromatography

The crude extract obtained in the previous step was partially purified using flash CC to obtain artemisinin rich fractions with less number of other components. 15 g of crude extract was separated on a 7 cm diameter column filled with normal phase silica. Adsorbent (silica gel) to solute (crude extract) ratio of 20:1 was used. Gradient type of elution was used to run the column under the applied pressure. Column was conditioned with 100% *n*-hexane and the gradient started with 100% *n*-hexane followed by 10% stepwise gradient from 100 % *n*-hexane to 100% ethyl acetate. Total of 54 fractions of 100 ml each were collected. All fractions were analyzed with thin layer chromatography (TLC) to identify the fractions containing artemisinin.

2.3. Dataset

The flash CC fractions 23 to 31 were found to contain artemisinin as confirmed by TLC analysis. These fractions were then analyzed by analytical HPLC on a Dionex UltiMate 3000 RSLC system equipped with a Diode Array Detector (DAD). ZORBAX Eclipse XDB-C18 reverse phase column (dimensions 150 × 4.6 mm (internal diameter), particle

size 5 μm) was used for separation. Eluent consisted of water and acetonitrile with 0.1% formic acid as modifier. Column temperature was adjusted to 35 °C. Sample injection volume of 10 μl and eluent flow rate of 0.8 ml/min was used. The original chromatograms obtained consisted of 19802 data points on retention time axis and 410 data points on wavenumber axis. In order to reduce the computational time every 10^{th} data point on retention time axis and every 5^{th} data point on wavelength axis was taken thereby reducing the size of matrix containing chromatograms for one sample at different wavelengths to 1981 × 82. Such kind of matrices of chromatograms for 9 samples were stacked one above another as shown in Fig. 1 to form a three way array of size 9 × 1981 × 82.

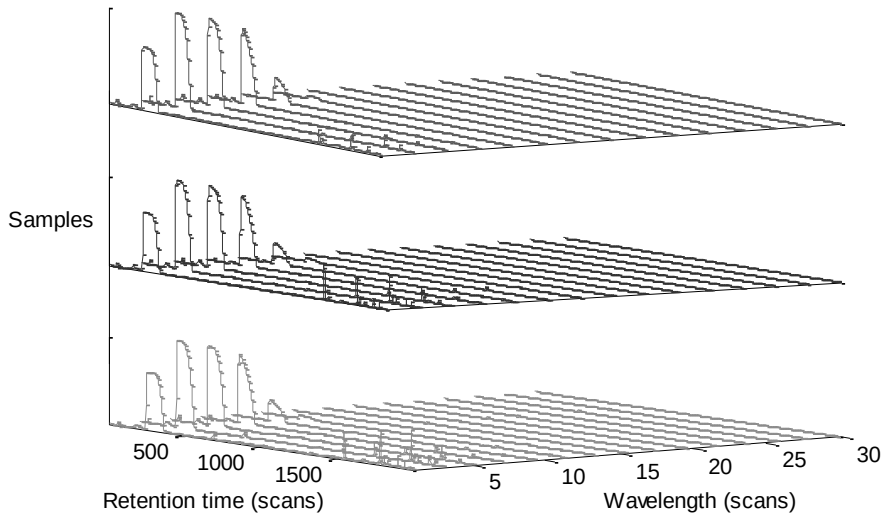

Figure 1. An exemplary three way data set of chromatograms. Green - Fraction 29; Red - Fraction 30; Blue - Fraction 31.

3. Results and discussion

3.1. Preprocessing of data

The dataset obtained after chromatogram measurements was subjected to preprocessing to remove the artifacts such as baseline drift, retention time shift and noise introduced by the fluctuations in the performance of instrument components and also due to the matrix effects (J. Amigo *et al*., 2010).

3.1.1. Baseline correction

Baseline drift is a commonly encountered problem during the measurement of chromatograms and it is important to remove baseline drift prior to the application of any multivariate data analysis technique. Practically it is difficult to avoid this problem during the measurements due to many parameters associated with it, therefore it is handled post measurement either by substracting the blank sample or substracting a polynomial fitted to the baseline points from the original chromatogram. In the present work, cubic spline algorithm was used to remove baseline drift which interpolates baseline fit with the help of selected datapoints on the original chromatogram and then applies to the dataset. The baseline correction algorithm was applied individually to

each sample matrix containing chromatograms measured at different wavelengths. Fig. 2a shows the chromatograms of fraction 29 measured at 82 different UV wavelengths before baseline correction and Fig. 2b shows the same chromatograms after baseline correction.

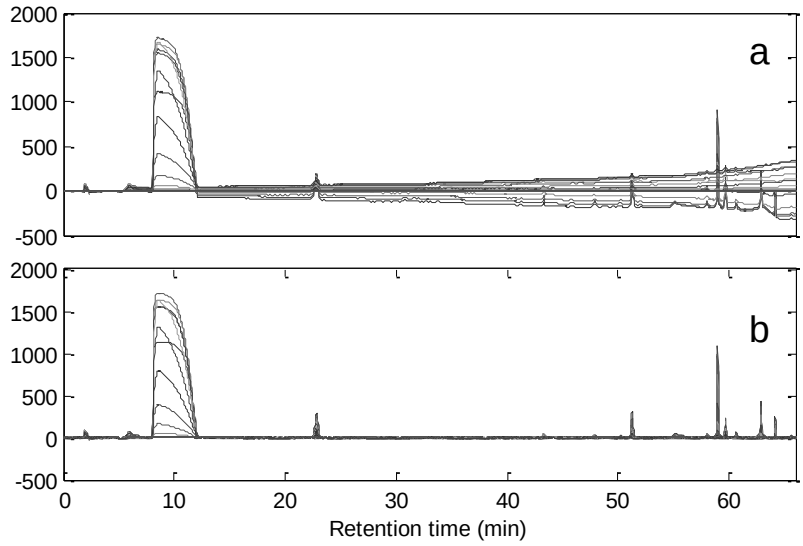

Figure 2. Exemplary chromatograms of fraction 29 measured at UV wavelengths from 190 to 600 nm. a) before baseline correction; b) after baseline correction.

3.1.2. Retention time shift alignment

Retention time alignment is an important pre-treatment step before the application of multivariate data analysis techniques especially in cases where identification and quantitation of peaks is required. In this work, the Interval Correlation Optimised Shifting algorithm (icoshift), which uses a piece-wise linear correction function based on an insertion/deletion (I/D) model and optimizes the piece-wise cross correlation using the Fast Fourier Transform was used to align the retention time shifts. The main advantage of using this method is that it does not rely on automatic peak picking procedures, but rather on basic segmentation procedures or on the experience of the analyst at identifying peaks or regions that are to be aligned (G. Tomasi *et al*., 2011).

3.2. PARAFAC modeling

In order to reduce the computation time, the preprocessed dataset was divided into total 10 intervals of retention times containing chemical signals to be analyzed as shown in Fig. 3. These intervals were selected in such a way that the signals from eluent solvents and background are excluded. PARAFAC was then applied to individual intervals to determine the total number of chemical components present in each interval, their relative concentrations, and pure UV spectra that can help in identifying chemical components. This information can be very useful in designing the downstream purification of artemisinin from these fractions by crystallization. Exemplary results from PARAFAC analysis of interval 8 containing artemisinin are shown in Fig. 4. One component unconstrained PARAFAC model was fitted to the raw data of interval 8 and it was confirmed that only one component model was enough to fit the data with the help of diagnostics such as explained variance, visualization of retention time loadings,

and residuals. Fitting of one component model indicates the presence of one chemical component. Results obtained from the PARAFAC model fitted to other intervals are shown in the Table 1.

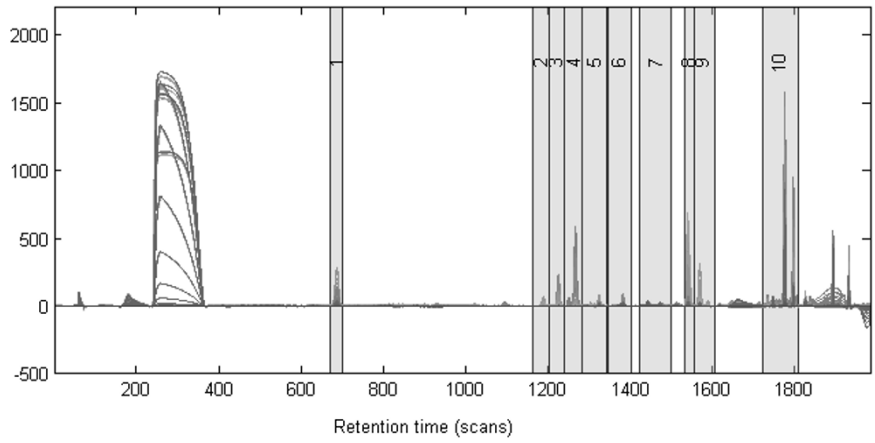

Figure 3. Intervals of dataset on retention time mode containing chemical signals for PARAFAC analysis.

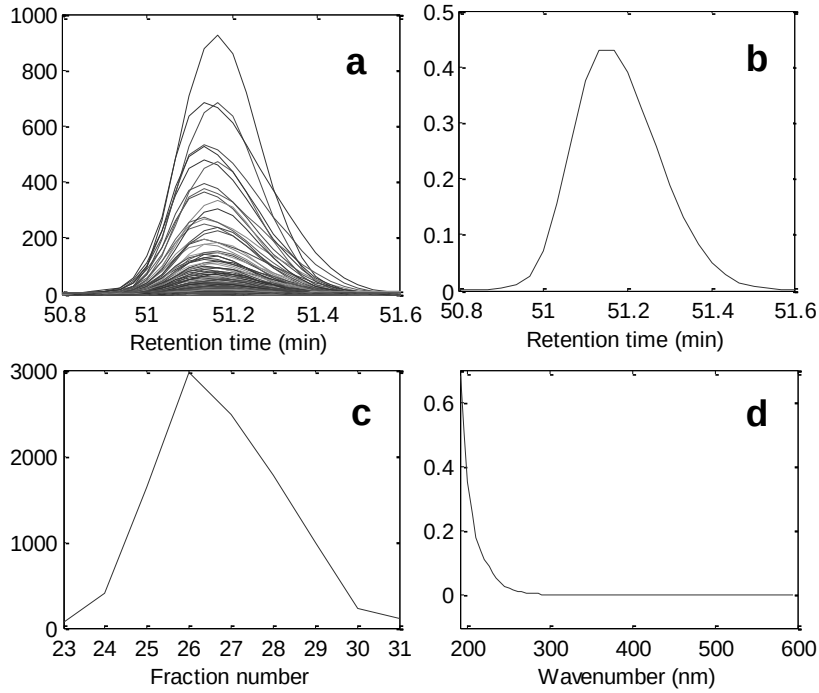

Figure 4. Results from one component PARAFAC model fitted to interval containing signal for artemisinin i.e. interval 8. a) Raw data; b) Retention time mode loadings obtained from the fitted model; c) Relative concentration profile of artemisinin; d) UV spectral mode loadings.

Table 1. Results obtained from PARAFAC model fitted to the individual intervals.

Interval No	Retention time range	No. of components	Explained variance (%)	Fractions containing these components
1	22.30 min – 23.30 min	1	98.778	27 - 31
2	38.63 min – 39.96 min	1	99.14	24 - 29
3	40.00 min – 41.13 min	1	98.825	23 - 25
4	41.16 min – 42.63 min	1	99.343	23 - 24
5	42.66 min – 44.63 min	2	95.11	23 - 26
6	44.66 min – 46.63 min	2	95.68	24 – 26 & 29 - 31
7	47.30 min – 49.83 min	2	93.442	23 - 31
8	50.63 min – 51.63 min	1	99.233	23 - 31
9	51.66 min – 53.30 min	1	94.638	23 - 31
10	57.16 min – 60.13 min	2	99.06	23 – 31 & 23 - 27

4. Conclusion

Multivariate data analysis technique PARAFAC was used to model the analytical chromatograms of fractions containing artemisinin obtained from flash CC. Application of PARAFAC for data analysis made the basic process information relatively easy to obtain such as number of chemical components present in the fractions along with artemisinin, their concentration profiles, and pure UV spectra, which is otherwise time consuming to do manually. In our future work, this technique will be used to analyze even more complicated data from hyphenated analytical techniques such as LC-MS and efforts will be made to identify the chemical components present in the fractions along with artemisinin.

References

J. Amigo, M. Popielarz, R. Callejón, M. Morales, A. Troncoso, M. Petersen, T. Toldam-Andersen, 2010, Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis, Journal of Chromatography A, 1217, 4422–4429.

R. Bro, 1997, PARAFAC. Tutorial and applications, Chemometrics and Intelligent laboratory Systems, 38, 149 –171.

P. Geladi, B. Sethson, J. Nystrfm, T. Lillhonga, T. Lestander, J. Burger, 2004, Chemometrics in spectroscopy Part 2. Examples, Spectrochimica Acta Part B, 59, 1347– 1357.

C. Malwade, B.-G. Rong, H. Qu, L. Christensen, 2012, Conceptual process synthesis for isolation and purification of natural products from plants – A case study of artemisinin from *Artemisia annua*, Computer - Aided Chemical Engineering, 1707-1711.

G. Tomasi, F. Savorani, S.B. Engelsen, 2011, icoshift: An effective tool for the alignment of chromatographic data, Journal of Chromatography A, 1218(43), 7832-7840.

S. Wold, 1995, Chemometrics; what do we mean with it, and what do we want from it?, Chemometrics and Intelligent Laboratory Systems, 30, 109-115.

U.S. Food and Drug Administration, 2004, PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.